Nucleotide sequence of a gene from chromosome 1D of wheat encoding a HMW-glutenin subunit

R.D.Thompson, D.Bartels and N.P.Harberd

Plant Breeding Institute, Maris Lane, Trumpington, Cambridge CB2 2LQ, UK

ABSTRACT

    A high molecular weight glutenin gene in hexaploid wheat has been
isolated by cloning in bacteriophage lambda and characterized. The gene
corresponds to polypeptide 12 encoded by chromosome 1D in the variety
"Chinese Spring". The coding sequence predicted contains seven cysteine
residues six of which flank a central repetitive region comprising more than
70% of the polypeptide. These findings are related to the role of high
molecular weight subunits in the viscoelastic theory of gluten structure.

INTRODUCTION

    The developing wheat endosperm is the site of synthesis and deposition
of a series of seed storage proteins some of which aggregate into a protein
complex known as gluten[1,2]. In wheat these storage proteins are
classified into two groups, the gliadins, which are soluble in aqueous
alcohol solutions, and the glutenins, which are alcohol-insoluble. The
glutenin fraction is made up of multimeric disulphide-linked aggregates
containing two size classes of polypeptides, the high-molecular-weight (HMW)
and low-molecular-weight (LMW) subunits. The LMW subunits, M_r 34-45 kD,
are encoded by genes at the Gli-1[3,4] loci on the short arms of the group 1
chromosomes of wheat. The HMW subunits, of fewer types, comprise
approximately 10% of the total glutenin aggregate and are encoded by the
Glu-1 loci on the long arms of the group 1 chromosomes[5]. They can be
distinguished from the gliadins and LMW subunits by their M_r of 70-90 kD
and their relatively higher glycine content (14-21 mole %)[6,7].
    The HMW subunits are believed to be largely responsible for conferring
the property of viscoelasticity on dough made with wheat flours[8]. This
property distinguishes wheat flour from that made with other cereals. The
Ewart model for gluten structure[9] suggests that formation of disulphide
bridges between cysteine residues of different subunits results in a network
of end-to-end polymers. It is known from amino-acid and DNA sequence data

that most of the small number of cysteine residues present in each HMW
subunit are located close to the subunit termini[10,11], although recent
evidence suggests that cysteine residues may also be present elsewhere in
the molecule[12]. Estimates of the accessibility of cysteine residues in HMW
subunits have indicated that at least some of them are involved in
disulphide bond formation[13]. That disulphide crossbridges are involved in
the formation of the elastic glutenin network is implied by the observation
that glutenin aggregates are broken down by reduction[6].

Comparisons of different wheat varieties have shown that there is
substantial intervarietal allelic variation between subunits. This
variation is controlled by three complex loci (Glu-Al, Glu-Bl, Glu-Dl).
Each locus encodes up to two different subunits. The pairs of subunits
contributed by the Glu-Bl and Glu-Dl loci are subdivided into x and y types
on the basis of their electrophoretic mobilities[14], NH_2-terminal amino acid
sequences[15] and chymotrypsin digestion patterns[16]. Hence Glu-Bl and Glu-Dl
are complex loci each controlling an x and y type subunit. Glu-Al controls
an x type subunit only in hexaploid wheat[16].

Variation between HMW subunits is known to affect the properties of the
glutenin aggregates containing them. Presence or absence of particular
allelic variants within wheat varieties is correlated with differences in
the bread-making quality of flour obtained from them. Comparisons of
complete polypeptide sequences of allelic HMW subunits may allow the
identification of structural variation associated with differences in
bread-making quality. Since these sequences cannot be easily obtained
directly from the HMW subunits themselves, cloning and sequencing of wheat
DNA fragments containing genes encoding HMW subunits has begun.

Using a characterized cDNA clone complementary to HMW subunit mRNAs[17],
chromosomal DNA clones of the Glu-1 loci have been isolated from a library
of wheat DNA fragments in λ-Charon 34. One clone selected from this library
has been shown to contain DNA from the Glu-Dl locus and to carry a gene
coding for the 1Dy HMW subunit of Chinese Spring.

MATERIALS AND METHODS
Strains and bacteriophages

E. coli strains ED 8800 (rk⁻mk⁻ SupE SupF lacZ M15 met⁻ RecA56) and
DH-1 (gyr A96 RecAl, endAl, thi-1 hsdR17 rk⁻mk⁺ SupE44) also called WL268,
were provided by Dr. N. Murray and Dr. W. Loenen respectively. K803 (SupE
met⁻ hsdS⁻ rk⁻mk⁻) was provided by Dr. N. Federoff.

The λ derived cloning vector λ-Charon 34[18] was provided by Dr. W. Loenen.

Library construction

High molecular weight wheat DNA was prepared from the variety Chinese Spring as described previously[17]. EcoRI partial digestion products of wheat DNA in the 15-20 kb size range (prepared on a sucrose gradient as described by Maniatis et al. (1982)[19], a gift of D.C. Baulcombe) were obtained. The λ-Charon 34[18] vector was prepared by digestion with BamHl and EcoRI followed by isopropanol precipitation[20]. The size-fractionated, wheat DNA EcoRI partial digestion products were ligated into the EcoRI sites of the λ-Charon 34 and λEMBL4 vectors and the mixture was packaged in vitro[19]. The packaged mixture was plated on E. coli strain K803 (RecA⁺).

Clone identification and purification

After the phage library containing wheat DNA was plated it was screened by plaque hybridization[21] using as probe the HMW glutenin cDNA pTag1290[17] labelled with ^{32}P by nick-translation. Hybridizing plaques were picked and purified by several rounds of plaque purification on E. coli strain WL268 (RecA⁻). λ clone DNAs were prepared as described in Maniatis et al.[19], following growth of phage on plate lysates. Cloned DNA inserts were subcloned in the pUC9 plasmid vector and grown on E. coli strain ED 8800 (RecA⁻).

RESULTS

Molecular cloning of wheat chromosomal DNA fragments containing genes encoding HMW-glutenin subunits

A wheat DNA library was constructed by ligating size-fractionated EcoRI partial digestion products of DNA of the variety Chinese Spring into the EcoRI sites of the cloning vector λ Charon 34[18]. The ligations were packaged in vitro and then plated on E. coli strain K803 (RecA⁺). Clones of interest were identified by hybridization to the insert of the HMW glutenin cDNA clone pTag1290[17] and then taken through several rounds of plaque purification using E. coli strain WL268 (RecA⁻) as host. Several λ-Charon 34 clones containing sequences hybridizing to the HMW glutenin cDNA were isolated and their DNA was purified for further analysis.

A previous publication[17] described the use of the nullisomic-tetrasomic lines of wheat to assign EcoRI fragments hybridizing to HMW glutenin cDNA to the chromosome from which they are derived. Similar analyses have enabled hybridizing fragments obtained by digestion with the restriction
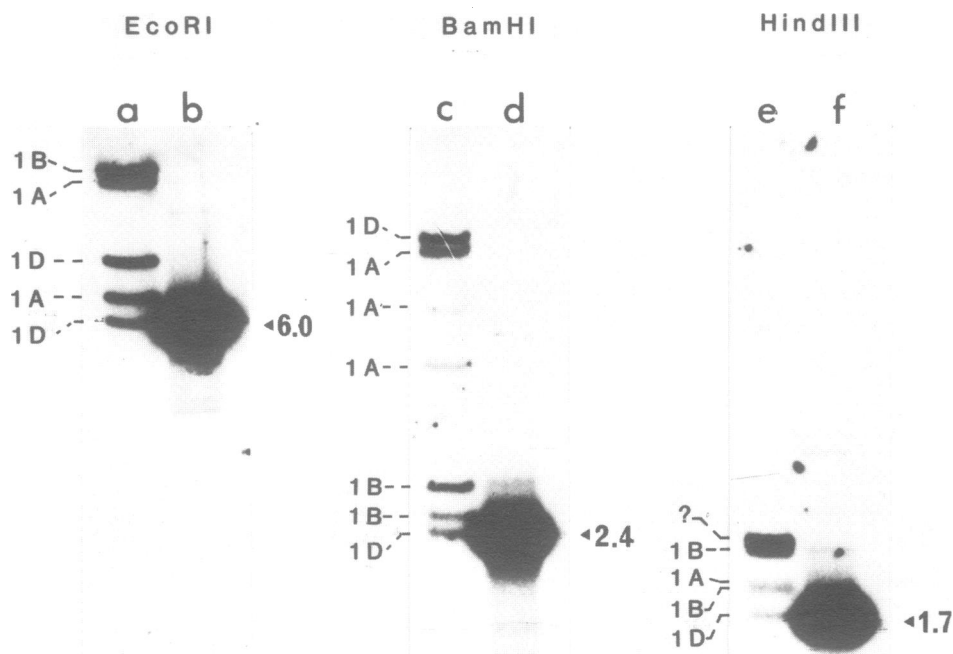
Figure 1
    Hybridization of the HMW glutenin cDNA (insert of pTag1290) to
(a) Chinese Spring wheat DNA digested with EcoRI, (b) λC11 DNA digested with
EcoRI, (c) Chinese Spring DNA digested with BamH1, (d) λC11 DNA digested
with BamH1, (e) Chinese Spring DNA digested with HindIII, (f) λC11 DNA digested
with HindIII. The chromosomal locations of the wheat DNA fragments are
indicated. Note that HindIII digests of Chinese Spring DNA also contain a
small (≈ 0.8 kb) fragment which hybridizes to the HMW-glutenin cDNA (Harberd
et al., in prep.). This fragment, derived from chromosome 1A, is not
visible in this experiment, because it has migrated too far. The sizes of
the hybridizing fragments derived from the λC11 clone are given in kilobase
pairs.

endonucleases BamH1 and HindIII to be assigned to chromosomes (Harberd et

al. in prep.). These data taken together allowed the integrity of putative

HMW glutenin genomic DNA clones to be confirmed (Fig. 1). The wheat DNA

restriction fragments hybridizing to the HMW glutenin cDNA were compared

with the fragments resulting from digestion of the cloned DNA with the same

restriction endonuclease. In the example in Figure 1 the cloned DNA digests

contain single hybridizing fragments which co-migrate with one of the

hybridizing fragments derived from the wheat chromosome 1D. This is

observed because there are BamH1 and HindIII sites outside the region of

hybridization to the cDNA and internal to the EcoRI sites at each end of the

cloned fragment (see Fig. 2). Hence λC11 is a clone of a 6.0 kb EcoRI
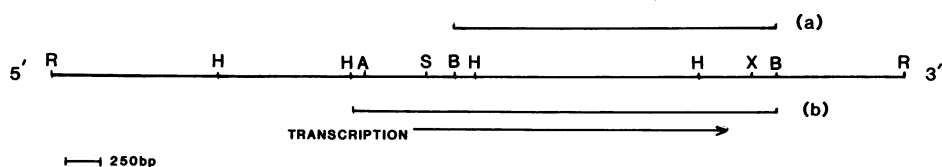
Figure 2

Map of the 6.0 kb EcoRI fragment in λCl1. (a) The BamHl fragment which hybridized to a HMW glutenin cDNA probe (pTag1290)[17]. (b) The sequenced portion. R = EcoRI, H = HindIII, A = AccI, S = SphI, B = BamHI, X = XbaI.

fragment containing HMW glutenin DNA sequence derived from chromosome 1D.

Gene localization and sequencing

The structure of the gene contained in the λCl1 clone was determined by restriction endonuclease site mapping and DNA sequencing. The region of the clone containing the HMW glutenin gene was localized on a restriction map of the entire cloned fragment (Fig. 2) by hybridization of the ^{32}P-labelled HMW glutenin cDNA sequence[17] to digests of the cloned DNA. Hybridization was observed to one BamHl fragment of size 2.4 kb. This BamHl fragment, which was also the only BamHl fragment to hybridize to a ^{32}P-end-labelled endosperm polyA+ RNA probe (prepared as previously described[22]), was taken and subcloned for sequencing. The DNA in this fragment was sequenced by generation of a series of Bal-31 deletions from both BamHl termini, and subcloning and sequencing of these deletions in M13 vectors[23,24]. Additional sequence data were obtained by dideoxy sequencing of HindIII and PstI restriction fragments, and from clones generated by sonication[25]. All clones were propagated on RecA⁻ E. coli hosts and under these conditions, no evidence for sequence instability was apparent. The open reading frame was found to extend beyond one of the BamHl sites and therefore the adjacent BamHl-HindIII fragment was also sequenced (see Fig. 2). The sequence was analysed using the Staden computer programmes[26,27].

Gene organisation

The mature amino-terminal protein sequences of several purified HMW-glutenin subunits have recently been determined[15]. These sequences, together with those from two cDNAs covering -COOH termini[11] enabled us to identify a unique open reading frame encoding a HMW glutenin polypeptide (Fig. 3). The coding sequence is uninterrupted by intron sequences and predicts a mature polypeptide of 68,617 M_r. The amino acid composition of the predicted polypeptide (eg. 35.5% Glutamine + Glutamic acid acid, 18% Glycine and 11% Proline[7,10]) matches that previously determined for HMW

```
ACCGCAGATTTGCAAAAGCAATGGCTAACAGACACCCAAACCCCAAGAAGCATTAACCACTTCTCTTAGATAAAAATAGCAGATTCGATA
     10        20        30        40        50        60        70        80        90


TACAAACGTCTACACTTCTGCAAACAATACCCAGAAGCCACAATTAGGATTGAACCGATTACGTGGCTTTAGCAGACCGTCCAAAAATCT
    100       110       120       130       140       150       160       170       180

                   -------------------
TTTTGCAAAGCTCCAATTGCTCCTTGCTTATCCAGCTTCTTTTGTGTTGGCAAATTGTTCTTTTCCAACCGACTTTATTCTTTTCACAT
    190       200       210       220       230       240       250       260       270

                                                                      -------
TTCTTCTTAGGGCTGAACTAACCTCGTCGTGCACAACCATTGTCCTGAACCTTGACCACGTCCCTATAAAAGCCCAACCAATCTCCACAA
    280       290       300       310       320       330       340       350       360

                                                                   M  A  K  R  L  V  L  F
TTTCATCATCACCCACAACACCGAGCAGCCACAAAATAGAGATCAATTCACTGACAGTCCACCGAGATGGCTAAGCGGCTGGTCCTCTTT
    370       380       390       400       410       420       430       440       450


 A  A  V  V  I  A  L  V  A  L  T  T  A  E  G  E  A  S  R  Q  L  Q  C  E  R  E  L  Q  E  S
GCGGCAGTAGTCATCGCCCTCGTGGCTCTCACCACCGCTGAAGGTGAGGCCTCTAGGCAACTGCAGTGTGAGCGCGAGCTCCAGGAGAGC
    460       470       480       490       500       510       520       530       540


 S  L  E  A  C  R  Q  V  V  D  Q  Q  L  A  G  R  L  P  W  S  T  G  L  Q  M  R  C  C  Q  Q
TCGCTTGAGGCATGCCGGCAGGTCGTGGACCAACAGTTGGCCGGTCGGCTGCCATGGAGCACGGGGCTCCAGATGCGGATGCTGCCAGCAG
    550       560       570       580       590       600       610       620       630


 L  R  D  V  S  A  K  C  R  S  V  A  V  S  Q  V  A  R  Q  Y  E  Q  T  V  V  P  P  K  G  G
CTCCGAGATGTTAGCGCCAAGTGCCGCTCTGTCGCCGTCAGCCAAGTCGCAAGACAATATGAGCAAACTGTGGTGCCGCCCAAGGGCGGA
    640       650       660       670       680       690       700       710       720

                                                                                   ==========
 S  F  Y  P  G  E  T  T  P  L  Q  Q  L  Q  Q  G  I  F  W  G  T  S  S  Q  T  V  Q  G  Y  Y
TCCTTCTACCCTGGTGAGACCACGCCACTGCAGCAACTCCAACAAGGAATATTTTGGGGAACATCTTCACAAACAGTACAAGGGTATTAC
    730       740       750       760       770       780       790       800       810

=======================> ==================================> ------------------ -----------
 P  S  V  T  S  P  R  Q  G  S  Y  Y  P  G  Q  A  S  P  Q  Q  P  G  Q  G  Q  Q  P  G  K  W
CCAAGCGTAACTTCTCCCTCGGCAGGGGTCATATTATCCAGGCCAAGCTTCTCCACAACAGCCAGGACAAGGGCAACAGCCTGGCAAATGG
    820       830       840       850       860       870       880       890       900

-----> -----------------> =========================> ------------------> -----------------> ==
 Q  E  P  G  Q  G  Q  Q  W  Y  Y  P  T  S  L  Q  Q  P  G  Q  G  Q  Q  I  G  K  G  K  Q  G
CAAGAACCAGGACAAGGGCAACAATGGTACTACCCAACTTCTCTGCAGCAGCCAGGACAAGGGCAACGATAGGAAAAGGGAAACAAGGG
    910       920       930       940       950       960       970       980       990

======================> ------------------> ------------------> ==========================> --
 Y  Y  P  T  S  L  Q  Q  P  G  Q  G  Q  Q  I  G  Q  G  Q  Q  G  Y  Y  P  T  S  P  Q  H  T
TACTACCCAACTTCTCTGCAGCAACCAGGACAAGGGCAACAAATAGGACAAGGGCAACAAGGGTACTACCCAACTTCTCCGCAGCACACA
   1000      1010      1020      1030      1040      1050      1060      1070      1080

--------------> ----------------> ------------------> ==
 G  Q  R  Q  Q  P  V  Q  G  Q  Q  I  G  Q  G  Q  Q  P  E  Q  G  Q  Q  P  G  Q  W  Q  Q  G
GGACAAAGGCAACAACCAGTACAAGGGCAACAAATAGGACAAGGGCAACAACCAGAACAAGGGCAACAACCAGGACAATGGCAACAAGGG
   1090      1100      1110      1120      1130      1140      1150      1160      1170

======================> ------------------> ------------------> ==========
 Y  Y  P  T  S  P  Q  Q  L  G  Q  G  Q  Q  P  G  Q  W  Q  Q  S  G  Q  G  Q  Q  G  H  Y  P
TACTATCCAACTTCTCCACAGCAGCTAGGACAAGGGCAACAACCAGGACAATGGCAACAATCAGGACAAGGGCAACAAGGGCACTACCCA
   1180      1190      1200      1210      1220      1230      1240      1250      1260

==============> ------------------> ==========================> ------------------> ==========
 T  S  L  Q  Q  P  G  Q  G  Q  Q  G  H  Y  L  A  S  Q  Q  Q  P  A  Q  G  Q  Q  G  H  Y  P
ACTTCTCTCAACAGCCAGGACAAGGGCAACAAGGGCATTACCTAGCTTCTCAGCAGCAGCCAGCACAAGGGCAACAAGGGCACTACCCA
   1270      1280      1290      1300      1310      1320      1330      1340      1350

==============> ------------------> ==========================> ------------------> ==========
 A  S  Q  Q  Q  P  G  Q  G  Q  Q  G  H  Y  P  A  S  Q  Q  Q  P  G  Q  G  Q  Q  G  H  Y  P
GCTTCTCAGCAGCAGCCAGGACAAGGGCAACAAGGGCACTACCCAGCTTCTCAGCAGCAGCCAGGACAAGGGCAACAAGGGCACTACCCA
   1360      1370      1380      1390      1400      1410      1420      1430      1440

==============> ------------------> ==========================> ------------------> ==========
 A  S  Q  Q  E  P  G  Q  G  Q  Q  G  Q  I  P  A  S  Q  Q  Q  P  G  Q  G  Q  Q  G  H  Y  P
GCTTCTCAGCAAGAGCCAGGACAAGGGCAACAAGGGCAAATCCCAGCTTCTCAGCAGCAGCCAGGACAAGGGCAACAAGGGCACTACCCA
   1450      1460      1470      1480      1490      1500      1510      1520      1530

==============> -----------------> ==========================> ------------------> -----> -----
 A  S  L  Q  Q  P  G  Q  G  Q  Q  G  H  Y  P  T  S  L  Q  Q  L  G  Q  G  Q  Q  I  G  Q  P  G  Q
GCTTCTCTCGCAGCAACCAGGACAACAAGGGCATTACCCAACTTCTCTACAGCAGCTAGGACAAGGGCAACAAATAGGACAGCCAGGACAA
   1540      1550      1560      1570      1580      1590      1600      1610      1620

------> ------------------> ------------------> ==========================> ------------------> ========
 K  Q  Q  P  G  Q  G  Q  Q  Q  T  G  Q  G  Q  Q  P  E  Q  E  Q  Q  P  G  Q  G  Q  Q  G  Y  Y
AAGCAACAACCAGGACAAGGGCAACAAACAGGACAAGGGCAACAGCCAGAACAAGAGCAACAACCAGGACAAGGGCAACAAGGATACTAT
   1630      1640      1650      1660      1670      1680      1690      1700      1710
```

```
==================> ----------------> ----------------> =========================> --------
 P  T  S  L  Q  Q  P  G  Q  G  Q  Q  Q  G  Q  G  Q  Q  G  Y  Y  P  T  S  L  Q  Q  P  G  Q
CCAACTTCTCTGCAGCAGCCAGGACAAGGACAACAGCAAGGACAAGGGCAACAAGGGTACTACCCAACTTCTCTCCAGCAGCCAGGACAA
    1720      1730      1740      1750      1760      1770      1780      1790      1800

--------> =========================> ----------------> ----------------> ---------------
 G  Q  Q  G  H  Y  P  A  S  L  Q  Q  P  G  Q  G  Q  G  Q  P  G  Q  R  Q  Q  P  G  Q  G  Q
GGGCAACAAGGGCACTACCCAGCTTCTCTGCAGCCAGGACAAGGACAGGGACAGCCAGGACAAAGGCAACAACCAGGACAAGGGCAA
    1810      1820      1830      1840      1850      1860      1870      1880      1890

--> ---------------> ----------------> =========================> ----------------> -----
 H  P  E  Q  G  Q  Q  P  G  Q  G  Q  Q  G  Y  Y  P  T  S  P  Q  Q  P  G  Q  G  Q  Q  L  G
CATCCAGAACAAGGGCAACAACCAGGACAAGGGCAACAAGGGTACTATCCAACTTCTCCACAGCAGCCAGGACAAGGGCAACAACTAGGA
    1900      1910      1920      1930      1940      1950      1960      1970      1980

-----------> =========================> ----------------> ==============
 Q  G  Q  Q  G  Y  Y  P  T  S  P  Q  Q  P  G  Q  G  Q  Q  P  G  Q  G.Q  Q  G  H  C  P  M
CAAGGGCAACAAGGGTACTACCCAACTTCTCCGCAGCAGCCAGGACAAGGGCAACAACCAGGACAAGGGCAACAAGGGCACTGCCCAATG
    1990      2000      2010      2020      2030      2040      2050      2060      2070

==========> ----------------> ----------------> ----------------> =====
 S  P  Q  Q  T  G  Q  A  Q  Q  L  G  Q  G  Q  Q.I  G  Q  V  Q  Q  P  G  Q  G  Q  Q  G  Y
TCCCCGCAGCAGACAGGACAAGCGCAACAACTAGGACAAGGCCAACAAATAGGACAAGTGCAACAACCAGGACAAGGGCAACAAGGGTAC
    2080      2090      2100      2110      2120      2130      2140      2150      2160

==================> ----------------> ----------------> ----------------> ---------------
 Y  P  T  S  L  Q  Q  P  G  Q  G  Q  Q  S  G  Q  G  Q  Q  S  G  Q  G  H  Q  P  G  Q  G  Q
TACCCAACTTCTCTGCAGCCTGGACAAGGGCAACAGTCAGGACAAGGGCAACAGTCAGGACAAGGACACCAACCAGGACAAGGGCAG
    2170      2180      2190      2200      2210      2220      2230      2240      2250

--> ---------------->
 Q  S  G  Q  E  K  Q  G  Y  D  S  P  Y  H  V  S  A  E  Q  Q  A  A  S  P  M  V  A  K  A  Q
CAATCAGGACAAGAGAAACAAGGCTACGACAGCCCATACCATGTTAGCGCAGAGCAGCAAGCGGCCAGCCCAATGGTGGCAAAGGCGCAG
    2260      2270      2280      2290      2300      2310      2320      2330      2340

 Q  P  A  T  Q  L  P  T  V  C  R  M  E  G  G  D  A  L  S  A  S  Q  *  *
CAGCCCGCGACACAGCTGCCGACAGTGTGTCGGATGGAGGGGGGCGACGCATTGTCGGCTAGCCAGTGATAGAACTCTCTGCAGCTTGCA
    2350      2360      2370      2380      2390      2400      2410      2420      2430

TGGTGCTTGGGCTAGCATGCACCTTATCTATCCAATAAACGTGATGCGTGTTCACAGTTTTTCATGTAACTAGAGTAAAACCCAATAATA
    2440      2450      2460      2470      2480      2490      2500      2510      2520

ATGCAAAATGAAAAGCTTCTCCAGCTAAAAAAGAACAAAATGGGTGTTGTATAGTATGGGTTGCATGTCTCAGCTCATTGTGTTGATGCA
    2530      2540      2550      2560      2570      2580      2590      2600      2610

TATTGATGACCCACAAGTATAGGGGATCAATTGTTGTCCTTTCGATAAGTAAGAGTAGTTGAATGCAACAAGGAGCAGAAGAAAATGATAA
    2620      2630      2640      2650      2660      2670      2680      2690      2700

GCGATTTTCAGCAAGGTATTTCTGCAAGCACTGGAAGTAACGGTAACATATAGTTTTGTCGTAGGATAATTCGTAGCAAGTAACAAGTAA
    2710      2720      2730      2740      2750      2760      2770      2780      2790

CAATGGTAGCAAGGTGCAGCAAGGTGGCCCAATCCTTTTTATAACAAGGGAAAAGCAGGAACGTTCTCTTATGTAAAGCAAAGTGCTCTC
    2800      2810      2820      2830      2840      2850      2860      2870      2880

GAGGACACATGAGAATTGTCATCTAGTCACGGTCATCATGTTTAGTTGATTTGTGTTTGCTACTTTGATAATTTGATATGTGGGTGGACT
    2890      2900      2910      2920      2930      2940      2950      2960      2970

GGTGCTTGGGTGATGTTCTTACTTGAACAATCAACCCACTGATGATTACCCCTTCTCACAAGCATCCGCAACTACGAAATAAGGATTAAG
    2980      2990      3000      3010      3020      3030      3040      3050      3060

ATAAATCTAACCATAGCATGAAACATATGGATCCCC
    3070      3080      3090
```

Figure 3

Sequence of a 3095 base pair region of the λCl1 insert. The DNA sequence is shown with a one-letter code translation of the reading frame utilized. Possible control elements are overlined. The site of polyadenylation of a related cDNA clone[11] is indicated (↓). Putative signal sequences for poly(A) addition are underlined. Arrowed bars indicate the positions of hexamer and nonomer repeat units within the coding sequence.
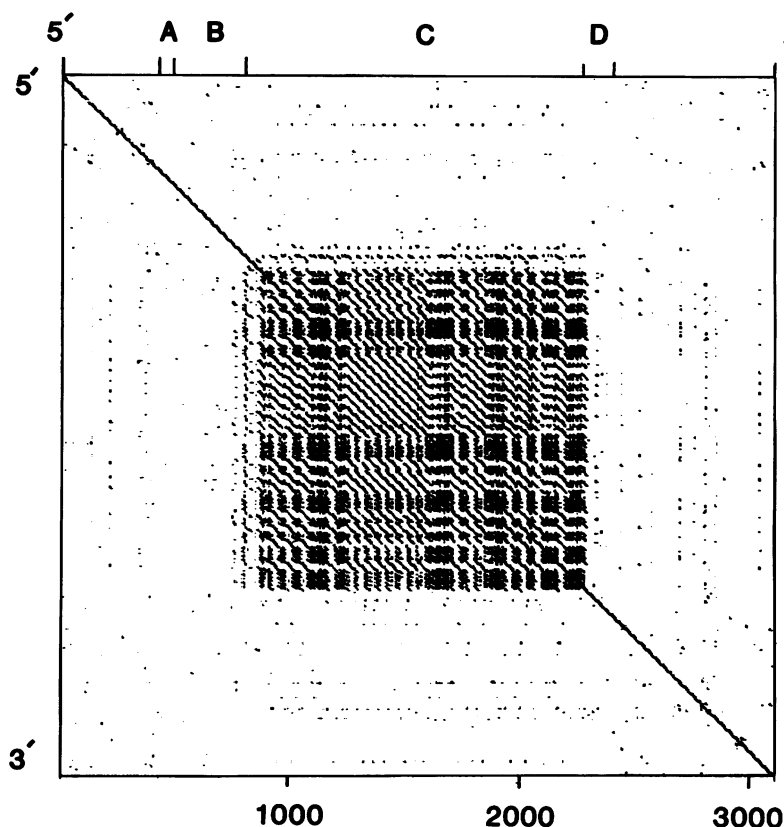
Figure 4

DIAGON[27] homology matrix of <u>Glu-D1</u> sequenced portion against itself. A dot is printed when 11 bases match in a 'window' of 15. Sequence domains A → D within the coding region are indicated. Segment A is the leader sequence, segment B is the non-repetitive amino-terminal portion, segment C is the repetitive region and segment D is the non-repetitive carboxyl terminal region.

glutenin subunits (32.6% Glutamine + Glutamic acid, 14.85% Glycine, 12.82% Proline). The mature NH_2-terminal sequence is preceded by a 21-residue leader sequence of characteristic amino acid composition[28].

The coding sequence of the mature polypeptide is similar to that of other prolamin storage proteins in that it can be divided into a number of distinct segments on the basis of amino acid composition. The HMW subunit has a tripartite structure, consisting of a non-repetitive amino terminal region, an extensive repetitive central region, and a non-repetitive

```
          (Highbury)
 1D2  (Copain)    E G E A S E Q L Q C E R E L Q E L Q E R E L K A C Q Q V P D(E)Q L Z D
          (Brigand)

 1D12 (Copain)    E G E A S R Q L Q C E R E L Q E    X Q L(K)A C(Q)X(Q)V

 1D
 Genomic  (Chinese   E G E A S R Q L Q C E R E L Q E    S S L E A C R Q V V D Q Q L A G
 Clone    spring)

 DNA-derived
 sequence
```

Figure 5

Comparison of amino-terminal protein sequences obtained for isolated polypeptides[15] with the mature amino-terminal protein sequence predicted from the gene characterized here.

carboxyl terminal region. The non-repetitive regions show homology to the corresponding non-repetitive regions of gliadin genes, and also some homology to sequences in certain globulin storage proteins of dicotyledonous plants. These homologies have been described in detail by Kreis et al.[29] The amino and carboxyl terminal regions contain most of the low abundance residues contained in the HMW subunit, including all but one of the cysteine residues (5 in the amino terminal region, 1 in the carboxyl terminal region).

The central region is composed of highly repetitive sequence, and is the only region showing extensive repetition in the DIAGON homology plot[27] (segment c in Fig. 4). The sequence consists almost entirely of two repetitive units, one a hexamer related to the amino acid sequence PGQGQQ, the other a nonomer related to the sequence GYYPTSLQQ. These units are interspersed such that single copies of the nonomer repeat separate segments containing several tandemly arranged copies of the hexamer repeat. Occasional length variants (eg. PGQQ, residues 353-356 of the mature polypeptide) of the basic hexamer and nonomer repetitive units are found in the sequence. Repeat units of both types display considerable variation at both the amino acid and DNA sequence levels. One nonomer repeat variant contains a cysteine residue (residue 525 of the mature polypeptide), probably a substitution for a tyrosine residue via a TAC→TGC codon change.

Comparison of the mature amino terminal sequences of a number of 1Dx and 1Dy HMW subunits with that of the polypeptide predicted by the gene sequence indicates that this gene encodes the 1Dy subunit of Chinese Spring known as subunit 12 (Fig. 5). The sequence is different in two respects to the 1D2 sequence; the Arginine for Glutamic Acid at residue 6 and a three codon deletion corresponding to residues 17-19.

```
                                    -210      Figure 6
    GLU      C C T T G C T T A T C C A G C T T      Comparison of sequence homology at
                                                    5' end of Glu-D1 sequence and
  ⍺-GLI      C C A T G C T T A T C T A G T T T      α-gliadin gene sequence[31].
                                    -387
                                   (-579)

                            -83
    GLU      C T A T A A A - A G C C C
  ⍺-GLI      C T A T A A A T A G C C C
                            -101
```

In Figure 3 420 bp of 5' non-coding sequence are also shown. This sequence contains a TATA box at -83 (numbered relative to the A of the starting ATG codon, Fig. 3). When the 5' untranslated region is compared, using the SEQH programme[30] to a 5'-untranslated region of an α-gliadin gene[31,32] a significant additional region of sequence homology is found (Fig. 6).

The 3'-untranslated sequence shows a high degree of homology to previously reported HMW glutenin cDNA sequences[11] and contains a putative polyadenylation signal at the same position as in those sequences, as well as two additional signals (Fig. 3) which may also be utilised, for example for longer mRNA species similar to that from which pC256[11] was derived.

DISCUSSION

(1) Clone stability

This paper describes the isolation, identification and sequencing of a cloned HMW glutenin gene from the Glu-D1 locus of Chinese Spring wheat. This clone was isolated from a wheat DNA library constructed using the λ-Charon 34 vector. Previous attempts to isolate HMW-glutenin gene clones from wheat libraries constructed using the λ EMBL 4 vector[20] and grown on a RecA⁺ host strain had been unsuccessful due to high levels of clone instability (data not shown). Use of the λ-Charon vector series for cloning wheat DNA has been described by others[33]. The λ-Charon 34-WL268 vector-host system is phenotypically RecA⁻RecBC⁻ and the resultant reduction in recombination presumably confers stability on otherwise unstable cloned DNA sequences[18]. However, the isolation of stable clones containing 13 different non-storage protein genes of wheat from libraries constructed in λEMBL series vectors and grown on K803 (RecA⁺) hosts (Baulcombe unpub.) suggests that different DNA sequences from the wheat genome are

differentially susceptible to instability during cloning.

The results presented in Fig. 1 demonstrate that no major rearrangements of the 6.0 kb EcoRI fragment occurred during cloning in the λ-charon 34 vector. It is of course not possible to rule out the occurrence of small scale rearrangements causing slight and hence undetectable alterations in DNA fragment mobility.

(2) Gene structure

The HMW glutenin gene cloned in λC11 possesses many of the features commonly found in DNA flanking the coding regions of eukaryotic genes including a TATA box, a translation start sequence and a polyadenylation signal which corresponds well to the canonical plant sequence[34]. There is however no sequence corresponding exactly to the CAAT or AGGA[34] boxes found in, for example, the α-gliadin genes[31,32]. The coding sequence of this gene, like that of other prolamin genes reported, is not interrupted by intervening sequences[34]. The reading frame identified in this gene is complete and is not broken by a nonsense codon such as is present in the predicted reading frame of the non-translated HMW-glutenin pseudogene described in the adjacent paper[41]. Also, there is good correspondence between the NH_2-terminal amino acid sequence determined from a purified 1Dy subunit[15] and the mature NH_2-terminal sequence of the polypeptide predicted by the gene sequence (Fig. 6). Two HMW glutenin mRNA species, one of 2700 bases and a second of 2200 bases, are specified by chromosome 1D[17]. The gene sequenced here is 2153 base pairs from the TATAAA box to the first possible polyadenylation site and therefore must specify the smaller RNA species.

Evidence from comparison of restriction fragments associated with the Glu-D1x and Glu-D1y alleles from several wheat varieties (Harberd, in prep.) suggests that these genes are present in single copies in the wheat genome. These considerations indicate that the gene contained in the λC11 clone is active at the level of transcription and translation and is the gene from the Glu-D1 locus encoding the 1Dy HMW-glutenin subunit of Chinese Spring known as subunit 12[8,12].

It is known that gliadins and glutenins are synthesised coordinately during endosperm development (Bartels, in prep.) and expression of the genes encoding them is therefore likely to be subject to a common regulatory mechanism. Sumner-Smith et al.[32] have noted the presence of a 17 bp direct repeat sequence present in two copies at -579 and -387 in the 5' untranslated region of three α-gliadin genes, and have suggested that this

sequence may be recognised by a developmentally regulated effector of gene expression. This sequence is part of a 56 bp sequence showing about 80% homology within the 5' untranslated region of the α-gliadin gene[31]. A sequence displaying strong homology to this repeat exists at -210 in the HMW-glutenin gene sequence described above (Fig. 3, 6) and in the sequence presented by Forde et al.[41]. These homologous regions may indeed therefore be involved in the regulation of storage protein gene expression. Equivalent sequences, sometimes termed enhancers, have been identified in the 5' non-coding regions of members of other multigene families[35,36]. In certain cases they have been shown to confer specificity of expression on coding sequences[37].

(3) HMW-glutenin polypeptides: structure and function

Wheat storage proteins are synthesised on membrane bound polysomes[38] and the α-gliadins have been shown to possess leader sequences[31,32]. The HMW-glutenin signal sequence lacks obvious homology with those reported for the gliadins but is of characteristic amino acid composition[28], possessing a lysine residue at position 2 in the sequence, followed by a stretch of hydrophobic residues and with alanine preceding the mature NH_2-terminal sequence.

The mature HMW-subunit contains three distinct regions, a non-repetitive NH_2-terminal region, a central region consisting of repetitive sequence and a non-repetitive -COOH terminal region. The sequences of the NH_2- and -COOH terminal non-repetitive regions display a high propensity towards formation of α-helix according to the rule for prediction of secondary structure from primary amino acid sequence[39]. The cysteine residues contained in these regions are likely to be available for disulphide bridge formation between or within the HMW subunits and other aggregating polypeptides. The central region, when tested for hydrophobicity using the parameters described by Chou and Fasman[39] is more hydrophobic than the non-repetitive NH_2- and -COOH terminal regions. Hydrophobic stretches are punctuated by several small hydrophilic pockets. The secondary structure of this region is a reflection of its unusual amino acid content. Computer predictions made from previous sequences of the region have suggested that tetraplets, which occur in both six-mer and nine-mer repeats, are involved in the formation of β-turn structures and that these multiply stacked β-turns give the molecule elastic properties[40].

The occurrence of a cysteine residue within the repetitive central region of this subunit is of interest. There are no cysteines in the

repetitive regions of the partial sequences reported previously[17], or in the central region of the gene sequence reported by Forde et al.[41]. The molecular environment of this cysteine is hydrophobic and this may reduce the probability of formation of disulphide bridges or affect the propensity for disulphide exchange. Whether this extra cysteine residue is involved in the viscoelasticity differences between proteins from different alleles can only be established by more comparisons.

The HMW subunits clearly possess many of the features predicted for them in the models which account for gluten viscoelasticity by inter-molecular end-to-end disulphide cross-linking connecting polypeptides with extensive elastic central regions in an elastic network[9]. The glutenin aggregate is a complex mixture of polypeptides and each of these presumably contribute in one way or another to gluten properties. The isolation of the genes for all the glutenin components including the HMW-subunit gene reported here, and LMW glutenin will enable a detailed assessment of the nature of the molecular interactions in gluten to be made. In particular, it will soon be possible to compare the sequences of allelic polypeptides known to differ in their effects on gluten properties and to characterise the structures associated with these differences.

REFERENCES
1. Kasarda, D.D., Bernardin, J.E. and Nimmo, C.C. (1976). In Advances in Cereal Science and Technology 1 (ed. Pomeranz, Y.) pp.158-236 (American Association of Cereal Chemists, St. Paul, Minnesota, 1976).
2. Payne, P.I., Holt, L.M., Lawrence, G.J. and Law, C.N. (1982). 'The genetics of gliadin and glutenin, the major storage proteins of the wheat endosperm'. Qual. Plant. Plant Foods Hum. Nutr. 31 (3) 229-249.
3. Jackson, E.A., Holt, L.M. and P.I. Payne (1983). Theor. Appl. Genet. 66, 29-37.
4. Jackson, E.A., Holt, L.M. and P.I. Payne (1985). Genet. Res. (in press).
5. Payne, P.I., Holt, L.M., Worland, A.J. and C.N. Law (1982). Theor. Appl. Genet. 63, 129-138.
6. Huebner, F.R., Donaldson, G.L. and J.S. Wall (1974). Cereal Chem. 51, 240-249.
7. Khan, K. and W. Bushuk (1979). Cereal Chem. 56, 505-512.

8. Payne, P.I., Harris, P.A., Law, C.N., Holt, L.M. and J.A. Blackman (1980). Ann. Technol. Agric. 29 (2), 309-320.
9. Ewart, J.A.D. (1977). J. Sci. Food Agric. 28, 191-199.
10. Field, J.M., Shewry, P.R., Miflin, B.J. and J.F. March (1982). Theor. Appl. Genet. 62, 329-336.
11. Forde, J., Forde, B.G., Fry, R., Kreis, M., Shewry, P.R. and B.J. Miflin (1983). FEBS Letters 162, 360-366.
12. Moonen, J.H ., Scheepstra, A.. and A. Graveland (1985). J. Cereal Sci. 3, 17-27.
13. Lawrence, G.J. and P.I. Payne (1984). J. Cereal Sci. 2, 225-239.
14. Payne, P.I., Holt, L.M. and C.N. Law (1981). Theor. Appl. Genet. 60, 229-236.
15. Shewry, P.R., Field, J.M., Faulks, A.J., Parmar, S., Miflin, B.J., Dietler, M.D., Lew, E.J-E. and D.D. Kasarda (1984). Biochim. Biophys. Acta 788, 23-34.
16. Payne, P.I., Holt, L.M., Thompson, R.D., Bartels, D., Harberd, N.P., Harris, P.A. and C.N. Law (1983). Proc. 6th International Wheat Genetics Symposium, Kyoto, Japan p.1125-1130.
17. Thompson, R.D., Bartels, D., Harberd, N.P. and R.B. Flavell (1983). Theor. Appl. Genet. 67, 87-96.
18. Loenen, W.A.M. and F.R. Blattner (1983). Gene 26, 171-179.
19. Maniatis, T., Fritsch, E.F. and J. Sambrook (eds) (1982). Molecular Cloning. Cold Spring Harbor Lab. New York.
20. Frischauf, A-M., Lehrach, H., Poustka, A. and N. Murray (1983). J. Mol. Biol. 170, 827-842.
21. Benton, W.D. and R.W. Davis (1977). Science 196, 180-182.
22. Bartels, D. and R.D. Thompson (1983). Nucl. Acids Res. 11, 2961-2977.
23. Poncz, M., Solowiejczyk, D., Ballantine, M., Schwartz, F. and S. Surrey (1982). Proc. Nat. Acad. Sci. USA 79, 4298-4302.
24. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.G.H. and B.A. Roe (1980). J. Mol. Biol. 143, 161-178.
25. Deininger, P.L. (1983). Analyt. Biochem. 129, 216-223.
26. Staden, R. (1982). Nucl. Acids Res. 10, 4731-4751.
27. Staden, R. (1982). Nucl. Acids Res. 10, 2951-2961.
28. Inouye, M. and S. Halegoua (1980). Crit. Rev. Biochem. 7, 339-371.
29. Kreis, M., Forde, B.G., Rahman, S., Miflin, B.J. and P.R. Shewry (1985). J. Mol. Biol. 183, 499-502.
30. Goad, W.B. and M.I. Kanefrisa (1982). Nucl. Acids Res. 10, 247-263.
31. Anderson, O.D., Litts, J.C., Gautier, M.F. and F.C. Greene (1984). Nucl. Acids. Res. 12, 8129-8145.
32. Sumner-Smith, M., Rafalski, J.A., Sugiyama, T., Stoll, M. and D. Soll (1985). Nucl. Acids Res. 13, 3905-3916.
33. Murray, M.G., Kennard, W.C., Droug, R.F. and J.L. Slightom (1984). Gene 30, 237-240.
34. Messing, J., Geraghty, D., Heidecker, G., Hu, N-T., Kridl, J. and I. Rubenstein (1983). From 'Genetic Engineering of Plants' eds. T. Kosuge, C.P. Meredith and A. Hollaender (Plenum Pub. Corp.).
35. Davidson, E.H., Jacobs, H.T. and R.J. Britten (1983). Nature 301,468-470.
36. North, G. (1984). Nature 312, 308-309.
37. Stuart, G.W., Searle, P.F., Chen, H.Y., Brinster, R.L. and R.D. Palmiter (1984). Proc. Nat. Acad. Sci. USA 81, 7318-7322.
38. Greene, F.C. (1981). Plant Physiol. 68, 778-783.
39. Chou, P.Y. and G.D. Fasman (1978). Ann. Rev. Biochem. 47, p.251.
40. Tatham, A.S., Shewry, P.R. and B.J. Miflin (1984). FEBS Lett. 177,205-208.
41. Forde, J., Malpica, J.M., Halford, N.G., Shewry, P.R., Anderson, O.D., Greene, F.C. and B.J. Miflin (1985). Nucleic Acids Research (accompanying paper).